DOCUMENT RESUME

ED 047 760 LI 002 642

AUTHOR	te Nuyl, Th. W.
TITLE	Examination of the Validity of the Conclusions Arrived at in the Aslib Cranfield Research Project.
INSTITUTION	Danish Centre for Documentation, Copenhagen.
SPONS AGENCY	International Federation for Documentation, The Hague (Netherlands). Committee on Classification Research.
REPORT NO	FID-CR-7
PUB DATE	68
NOTE	29p.; FID Publ. Serie-No. 405
AVAILABLE FROM	FID/CR, Danmarks Tekniske Bibliotek, Oster Voldgade 10, Copenhagen K., Denmark ($1.50)
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Indexing, *Information Retrieval, *Library Research, Performance, *Relevance (Information Retrieval), Reliability
IDENTIFIERS	*Cranfield Project

ABSTRACT

The Cranfield Project was done in a laboratory sphere. In a practical situation, it is impossible to determine how many documents are relevant to a certain question. Although in the Cranfield Project collections of different sizes have been used, these collections do not fulfill the requirements that the small collections are randomly taken from the larger one. On the contrary, all the relevant documents in the small collection are the same as those in the large collection. The research work done does not give information to what extent the sampling method gives practical results. A more attractive method would be to accept figures obtained in the Cranfield Project for recall ratios at various coordination levels. In the Cranfield Project, coordination level of 3 means that out of 7-10 search terms all possible combinations are taken, but are counted only once independent of which terms match. This is inherent to the scan-column index technique used in the project, which is not easy to imitate in operational systems using computers, machine punched cards or manual punched cards. (Author/MF)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

FID Publ. Serie-No. 405 UDC 025.4

ED 047760

International Federation for Documentation
Committee on Classification Research

FID/CR Report Series
Report No. 7

EXAMINATION OF THE VALIDITY OF THE CONCLUSIONS ARRIVED AT IN THE ASLIB CRANFIELD RESEARCH PROJECT

A study made on behalf of the FID/CR
Committee on Classification Research

by

Th.W. te Nuyl

LI 002 642

Danish Centre for Documentation
Copenhagen 1968

ZEIST, NEDERLAND, 1967

FID/CR
Nov. 1968

EDITORIAL

The publication of this issue of the FID/CR Report Series has unfortunately been delayed by the activities of the Secretariat in arranging the *Seminar on UDC in a Mechanized Retrieval System,* Copenhagen, 2—6 September 1968. It makes available for the readers two papers which for some time have been known only within smaller groups.

Report No. 7 by Th. W. te Nuyl, former chief of the Patent Documentation Department of the Dutch Shell Company, The Hague, is a study made on behalf of the FID/CR with financial support from the FID Committee budget. The CR-Committee is greatly indebted to the author for this critical extract of the two volumes, published by Cyril Cleverdon, Jack Mills and Michael Keen on the *Factors Determining the Performance of Indexing Systems,* with special reference to the contents of the conclusions. Various suggestions for further research, based on the rich material presented in the Cranfield reports, are set forth, and to some extent later followed up by the author.

Report No. 8 by Richard S. Angell, chief of the Technical Processes Research Office, Library of Congress, Washington, D.C., contains *Two Papers on Thesaurus Construction* which date back to the Tokyo Conference in 1967. The papers have been re-edited to make a whole, and include a number of comments received from the editors of the 10 thesauri in question. The relational mechanisms of these thesauri have been studied in detail with the aim of presenting a kind of analysis that may have value in classifying the language and structure of subject access vocabularies with a view to achieving maximum compatibility among them.

It is noted that the FID/CR Report Series is being met with a growing interest as a useful means of communication between the Committee and documentalists from all parts of the world. Just because of its cheapness a considerable part of the edition is distributed without charge; this applies especially to requests received from countries with valutary difficulties.

R. M. H.

PREFACE

The Cranfield reports contain a wealth of information based on very hard and devoted work. One would like to get much use out of them.

When looking, however, at the conclusions, these appear to be limited and to be bound to the particular environment of the test.

A close study of the reports can reveal the reasons. The scientific methods have the boundaries set by science, and information cannot be fully understood within these boundaries. After discovery of this fact it is no longer possible to look upon the data contained in the reports with the expectation to find a complete answer, but, in trying to understand the presentation of the data the attitude will prevail to find out what is missing or which factors have been overlooked.

From this the reader of this report should not draw the conclusion that the Cranfield project has not been of great importance. The fact that the work done at Cranfield has been so elaborate gives the possibility of getting a deep insight in the matter. This report does not deal with all the subjects discussed. There is room for further intensive study of the matter. In particular when the conclusion of this report, that there is no complete answer to the problem of comparing and evaluating information systems, is accepted, the way will be open to find out under what conditions the best comparison of systems can be made and how systems may best be evaluated.

For philosophers the reports must be most welcome, for they provide the material to widen the horizon of our present materialistic way of thinking and to approach the real nature of thought. In this way they will enable to accept a metaphysical or spiritual world and build a bridge between science and religion.

For computer experts the possibility is now given to more intelligently understand what the computer is doing. That it is not thinking itself, but performing under fixed conditions only certain aspects of thinking.

The mathematician may find the impulse to develop mathematics for higher class logics to master the new field opened by information.

The documentalist will on the basis of the Cranfield reports be in a position to better understand his task, which will require a highly developed thinking freed from the limitations set by natural science and mathematics.

INTRODUCTION

The results of the investigation into the factors determining the performance of indexing systems, known as the ASLIB Cranfield Research Project and supported by a grant to ASLIB by the National Science Foundation, have been published in two volumes.

The first volume on the design of the project by Cyril Cleverdon, Jack Mills and Michael Keen has been published in 1966 in two parts; part 1 with text and part 2 with appendices.

The second volume is on test results and has been written by Cyril Cleverdon and Michael Keen.

The main conclusion that arises from the project is that the single term index languages are superior to any other type. The authors of the report are of the opinion that this conclusion seems to offend against every canon on which they were trained as librarians and that it is bound to throw considerable doubt on the methods which have been used to obtain the results.

The Classification Research Group of the International Federation Documentation (FID/CR) has undertaken to examine the validity of the conclusions and this report is intended to be an introductory examination for a further discussion of the matter in FID/CR.

The methods used in the Cranfield Project have been fully discussed in the reports. However, to make the present presentation complete in itself, it will be necessary, in addition to an exposition of the Cranfield Project itself, to repeat here the description of the methods. This will be done as far as possible with the exact wording of the Cranfield reports in order to avoid any distortion of meaning.

The ASLIB Cranfield Research Project

In the ASLIB Cranfield Research Project there are two stages.

The first stage commenced in July 1957, when the grant by the National Science Foundation was awarded, which made it possible to undertake actual work at the College of Aeronautics, Cranfield, England, on proposals submitted in the year before. It was directed towards an investigation into the comparative efficiency of indexing systems. Four systems were tested; one in which natural language was used, another based on a controlled vocabulary, a further system with alphabetical arrangement and one with classified arrangement. The "concept indexing" - defined further on - was in general the same for all four systems.

The original ASLIB-Cranfield investigation did not, by itself, produce firm answers to what is one of the basic problems in information retrieval, namely the decision as to which language should be used. The four systems investigated are considered to deal with four index-languages.

In regard to the second stage the main importance of the earlier project was in the new hypotheses which could be formulated. Some of these are:

(8) The most important factors to be measured in the evaluation of information retrieval systems are recall and precision.

(10) The index language has a relatively minor effect on the operational performance of an information system.

(11) Given the same concept-indexing, any two or more kinds of index-languages will be potentially capable of similar performance in regard to recall and precision.

(12) The more complex an index-language, i.e. the more devices it incorporates, the greater the range of performance in regard to recall and precision.

The second stage started in 1962 and was also made possible by a grant from the National Science Foundation. It was concerned with the said basic problem in information retrieval, namely which index-language should be used.

In Cranfield I, an additional tool had been provided for all four systems investigated, through an equally effective "lead-in-vocabulary", by which term is implied a complete list of all the sought terms.

There are distinguished three types of terms:
lead-in terms, representing concepts described by other terms,
code-terms , which are actually used in indexing, and
index-terms , which are in addition to code-terms, any combination thereof which make up and express new concepts.

There are many devices which are used in various ways to make up different index-languages. Any basic research on index-languages with the purpose to advance knowledge of information retrieval requires, according to the authors, a laboratory-type situation, where the performance of index-languages can be studied in isolation.

Since relevance assessments were known to cause problems, one limitation was essential in the design; actual questions should only involve relevance assessments by the questioners.

The technique adopted was using prepared questions based on source documents and the first objective was the precise measurement of recall and precision ratios. The prerequisite is the determination of the sets of documents which are and are not relevant to each of a set of test questions. The point of view taken is, that "with the aid of the set of documents and the set of questions (for which the document/question relevance assessments have been previously made by the questioners) it will be possible to test each index language device in turn and so get precise figures for the effect on recall and precision ratios".

For the project it was considered that the nearest to the ideal would be a combination of using actual questions that have been put to an I.R. system, with a relevance assessment being made by the questioner, who would be a scientist.

In respect of the determination of the recall ratio, there was only one way, namely to look at every document in relation to every question. This places a restriction on the size of the test collection and the number of questions to be searched. There seemed to be some advantage in having a large number of questions in relation to the number of documents in the collection and the aim was 1,200 documents with 300 questions.

The method adopted to obtain the documents and the questions was to select a number of recently published research papers, The author of each paper was to be requested to provide the basic problem, in the form of a question, and also to give some additional problems which had arisen in the course of his work. At the same time he would be asked to state which papers in his list of references were relevant to the various questions he had provided. It was intended that the document collection would be made up of the papers that had been included as references.

In order to obtain the cross-check of every document and every question, first a screening process was done by postgraduate students to eliminate most of the non-relevant documents for each question. Those papers which had a reasonable possibility of being relevant were sent to each author to make a final decision concerning relevance.

The investigation was based on the belief that all index languages are amalgams of different kinds of devices. Such devices fall into the two groups of those which are intended to improve the recall ratio and those which are intended to improve the precision ratio.

The first stage for the storage of the information was to determine and fix the "concept indexing" of the documents and the relationships of the concepts. By "concept indexing" is meant the decision as to which concepts and groups of concepts are significant from the viewpoint of retrieval. Such concept indexing can only be in the terminology of the document. As soon as there is any translation of the document terminology to any kind of formalized language, then one of the index language devices must have been brought into use.

The indexing technique and also the searching method permitted matching any actual word in the question with any term used in the concept indexing and then to introduce all the devices by stages.

Further it was clear that if there was to be any comparison of

experimental results, it was necessary first to investigate the effect on performance of the generality ratio, namely the relationship between the number of relevant documents and the size of the collection. It was planned to measure the effect of this factor on recall and precision.

Documents and questions were obtained from about 200 authors of research papers and in the first letter to these, it was stated, that it was the intention to attempt to refine the measurements of performance and reach a point where systems can be designed to meet given performance and economic requirements.

In a second letter to authors information was asked in respect of:

Relevance assessments of documents in the total collection of 1400 documents,

Associative indexing, that is determining the most useful association of terms to be used in searching,

Citation indexing,

Weighting of concepts,

Related search terms, and

Rephrasing of original question.

The assessment of relevance was to be based on the following scale of five definitions:

1. References which are a complete answer to the question.
2. References of a high degree of relevance.
3. References which were useful.
4. References of minimum interest.
5. References of no interest.

Of the original 640 questions received from authors, 361 were selected, that had two or more documents assessed as relevance grade 1, 2 or 3 and that were grammatically complete.

For 86 of the 361 questions no other documents were considered to be relevant. For the other 275 questions there was at least found one document judged as possibly relevant. When submitting these documents to the authors there were added in total 198 extra documents (resulting from a test of the questions by the technique known as bibliographic coupling), that had seven or more references in common with one of the author's cited relevant papers of grade 1, 2 or 3.

In addition to making relevance assessments, the authors were

asked to indicate the relative importance of each term or concept in the question by marking with a "weight" from the following scale:

1. A paper that did not cover this term would be of no use,
2. It is desirable that this term should be covered by the document,
3. This is a term which is not absolutely essential to the inquiry.

As a final result 279 questions of the original 641 questions were available for test. Of these 279 questions 118 are basic and had given rise to the research work. The remaining 161 questions are supplementary.

For these 279 questions 1961 documents were accepted as relevant, which means an average of 7.0 documents per question.

Of the 1961 documents 171 (0.6), 461 (1.7), 902 (3.2) and 427 (1.5) were considered to have respectively the grades of relevance 1, 2, 3, and 4. The average for each question appears in brackets.

The questions mainly fall into two areas: high speed aerodynamics and aircraft structures. They varied in length; the search terms ranged from 2 to 15.

The "source" document for each question is removed from the collection when that question is being tested and does not appear in any of the results at all.

INDEXING

The simplest known indexing device was taken to be that of condensation of the full text into an index language consisting solely of the "uniterms" thrown up by the title and text of the document itself.

The indexing has to be exhaustive, to be able to test recall devices. Exhaustivity in indexing refers to the degree to which one recognizes (includes) the different concepts or notions dealt with in a document.

The indexing has to be specific to be able to test precision devices. Specificity refers to the generic level at which concepts or notions are included in the indexing.

Devices which increase recall are:

confounding synonyms;
confounding word forms;
generic-hierarchical linkage (genus/species relationship);

non-generic hierarchical linkage, which is essentially one of conjunction ("coordination"), rather than complete inclusion, and deals with particular relations, e.g. between a thing and its parts, a thing and its properties.

There are other devices, which have not been included in the project, like:
bibliographic coupling;
associative indexing, which is based on frequency of occurrence and co-occurrence of individual words and of particular word clusters;
grouping of words, to reduce the vocabulary, e.g. by dictionary-based clusters.

Devices which increase precision are:

coordination - i.e. the conjunction of two or more terms, and may be precoordination and postcoordination;

weighting - i.e. the assignment to a term of a figure representing the relative significance of that term in the total subject description of the document;

links - i.e. indicating a particular connection between two or more terms;

roles - i.e. indicating the role or function of a particular term in an indexing description.

The adopted way of indexing would be in uniterms (postcoordinate) and take into account the precision devices of weighting, links and roles.

The recall devices of synonyms, word-forms and hierarchical linkage (generic and non-generic) would be measured by variations in search programming.

The documents were analysed by the indexers in four stages. Firstly, concepts were distinguished. Secondly, the concepts were grouped to display "themes" into which the document could be partitioned. Thirdly, six different weights were given to the concepts and subsequently the same weights to the terms. If these appeared in more than one concept the weighting of the more heavily weighted concept was given. Fourthly, the terms with their weights were written out.

The preparations for the evaluation of roles, to be assigned to the terms in a fifth step in the indexing were temporarily abandoned.

As a further device to enlarge classes (=document groups based on

similarities) there is reference to the use of quasi-synonyms. These are terms which can be used synonymously in certain contexts, but which are not true synonyms. The continued separate use of the terms confounded is not excluded and no figure showing the exact degree of vocabulary reduction is possible.

Vocabulary size was at a certain stage considered to be the main determinant of recall and precision and the first testing of hierarchy took the form of a fixed reduction in vocabulary size.

Reduction of classes hierarchically should take note of the weight of litterature in the different classes. This led to the noting of word frequencies in determining which classes should be retained intact at a particular level of reduction.

Although the indexing was done for 1400 documents, only a subset of some 200 documents, containing all the documents relevant to some 40 questions, was used for the preparation of "concept" languages. In order to make the new collection reasonably homogeneous only aerodynamics documents were included.

The "concept" schedules, Appendix 5.4, were essentially "one-place" schedules in linear sequence, and the function was simply to show the hierarchical relations (generic and non-generic) between the terms (concepts).

Connections between classes or concepts may be shown by multiple entry in a classified index or by a rotated A/Z relative index, in which all the different contexts in which a term appears are gathered together as qualifiers of that term. Each concept appeared as many times as it had distinct words.

A second subset of 350 documents was selected, which included the first subset of 200 documents. The indexing of these documents was compared with a conventional index, the Engineers' Joint Council Thesaurus of engineering terms. An extension thereof was, however, necessary if the specificity were not to suffer seriously. As a result of the different connectives in the E.J.C. there was a different way of recognizing synonyms and other connectives between terms and between concepts. The rejected terms and phrases constructed a massive "lead-in" vocabulary from the terms and expressions of the natural language to those of the E.J.C. languages. Over 1500 entries were made for the subset.

TESTING TECHNIQUE

To validate the proposed design of the tests it was decided to make a small test based on 116 documents and 14 questions, for which there were 26 relevant documents. It was decided to investigate 5 sets of recall devices and 4 sets of precision devices, based on the single term natural language indexing.

Each recall device, or in other words, each index language, gives rise to a separate system in which the precision devices of coordination, weighting, links and roles can be investigated, requiring, however, many different searches.

Manual re-punching of new indexes, as would be required when using a peek-a-boo system, would have been a big task. Also the application of computers did not promise a solution. Finally the "scan-column" index, which later on appeared to be known and used in the Dutch Patent Office, was adopted.

For this purpose, finally 361 sets of search sheets were made, 23 sheets in each set showing all the 1400 document code numbers and posted with all the occurrences of the terms to be used in searching each question. There were in fact 361 question-indexes, showing for each question the documents having search terms, as also which ones, with their synonyms, word endings and quasi-synonyms, indicated by means of code letters. Further the weights were indicated.

The 361 questions which it was proposed to use for searching produced a total of 723 different terms and these became known as starting terms. As such they were terms used in the questions without being subjected to any controls and were equivalent to the natural language index terms. The starting point of each series of tests is the use of the basic terms as indexed.

The first series of searches were performed on single terms. Three variables were investigated:

1. six index languages;
2. simple coordination;
3. the three levels of indexing exhaustivity, indicated by the weights 5-6 for concepts in minor subsidiary theme,
   7-8 for concepts in major subsidiary theme, and
   9-10 for main general theme.

The results were scored on a sheet, showing the document numbers on the left hand side and across the sheet figures indicating the number

of search terms that match with the document terms and thereby the co-ordination levels, and that for each of the six languages at each of the three levels of exhaustivity, in fact the levels 5-10, 7-10 and 9-10.

Since any combination of search terms was to be accepted, it was not necessary to note which search terms occurred. Therefore, combinations which were accepted could be without sense. In later searches such unwanted combinations were eliminated.

The final results for a question were recorded on a results-sheet, showing three separate tables for the three levels of exhaustivity, each table recording for a number of coordination levels the score of relevant and non-relevant documents for the six languages. For the co-ordination levels 1+ and 2+ no non-relevant figures are given.

The figures obtained for the various questions are totalled to provide results for a set of questions.

The investigation of the precision devices of interfixing and partitioning was done by examining the original indexing sheets for the relevant and non-relevant documents that had been retrieved as a result of the searches.

For the testing of the simple concepts 16 aggregates of recall devices were tested.

The results obtained in the preliminary tests were very similar to those obtained with the complete collection (when due allowance is made for the generality ratio). The best performance was obtained with the group of eight index languages which used single terms. The group of fifteen index languages which were based on concepts gave the worst performance, while a group of six index languages based on the Thesaurus of Engineering Terms of the Engineers ' Joint Council were intermediary. Of the single term index languages, the only method of improving performance was to group synonyms and word forms, an any broader grouping of terms depressed performance. The use of precision devices such as links gave no advantage as compared to the basic device of simple coordination.

TEST RESULTS

In Volume 2 the test results obtained with the complete collection have been reported. Chapter 2 presents the various factors of the tests,

showing changes regarding details in comparison with the original plan. The main points may be summarised.

The testing also covered searches which accepted any single term in the question.

On the single term languages besides coordination also partitioning and interfixing were tested.

On the controlled terms weighting was tested. Weights were assigned to the search terms and a match sought with the weights assigned to the terms in indexing.

A new point discussed in chapter 2 concerns the application of search rules, which concerned the combinations of terms that were accepted.

The most satisfactory and carefully applied search rules were applied to controlled language tests, since it was thought that intelligence in searching would be best tested on an index language that also had an average degree of intelligence used in its formulation. This was the search rule E, where all the combinations of acceptable terms were individually selected for each coordination level and it was applied with and without the precision device of weighting.

One further additional rule was to make a record of the number of starting terms that came up in a given match, to distinguish from a match with any related term.

The results of the tests can be recorded in a composite table showing the numbers of retrieved documents for the various combinations of variables. These are indicated by roman numerals, Arabic numerals, capitals and lower case letters. The meaning of the various characters is as follows:I - single terms; II - simple concepts; III - controlled terms; the possible 8 single term index languages are indicated by the I, followed by a 1-9, with the 4 missing:

I-1 - natural language,
I-2 - natural language + synonyms,
I-3 - natural language + word forms,
I-5 - natural language + synonyms + quasi-synonyms,
I.6 - natural language + word forms + synonyms + quasi-synonyms,
I.7 - natural language + syh. + first hierarchical reduction,
I.8 - natural language + syn. + first + second hier. reduction,
I.9 - natural language + syn. + first + second + third.hier.red.

For the simple concept indexing there were 15 languages, indicated by II, followed by a 1-15, and for the controlled term indexing there were six languages coded by III, followed by 1-6.

The precision devices are coded a, b, c, d and e:

a - coordination,
b - a + partitioning,
c - a + interfixing,
d - a + partitioning + interfixing,
e - a + weighting

The search rules are coded A, B, C, D, E and F:

A - any combination of terms accepted,
B - single terms grouped into concepts, and subordinate terms not accepted without this basic term,
C - selection of terms made from original question, any combination of the selected terms accepted,
D - specific combinations of the selected terms demanded,
E - sets of specific combinations demanded at each coordination level,
F - matching demand in terms of language 1.

The coordination level is indicated by an Arabic numeral, sometimes followed by a "+" sign to indicate that all higher levels are included.

Document relevance is indicated by a 1, 1-2, 1-3 and 1-4.

Exhaustivity is indicated by 1, 2 and 3. Exhaustivity 1 is the least exhaustive indexing where the average was 13 terms per document. With exhaustivity 2 the average was 25 terms per document and with exhaustivity 3 is meant the full indexing which averaged 31 terms for each document. In tests based on titles the average number of terms was 7 and for abstracts this number was approximately 60.

METHODS FOR PRESENTATION OF RESULTS

The measures have to reflect the changes in the particular component being tested. In addition it should be possible to make comparison between different sets of test results.

The main component to be tested was a range of index languages. The measures are numbers of documents, relevant and non-relevant, retrieved and not-retrieved, which are obtained by searching at different coordination levels.

By means of the numbers different aspects of performance can be expressed. This has been done as follows:

a. by denoting the documents which are retrieved and relevant;
b. which are retrieved and non-relevant;
c. which are not-retrieved and relevant, and
d. which are not-retrieved and non-relevant; the single performance measures that can be used can be listed as follows:

$a : (a + c)$ = recall ratio, the fraction of relevant documents retrieved,
$a : (a + b)$ = precision ratio, the fraction of the retrieved documents which are relevant,
$c : (a + c)$ = the fraction of relevant documents that are missed,
$b : (a + b)$ = the fraction of retrieved documents that is non-relevant,
$b : (b + d)$ which is called fall-out, and
$d : (b + d)$ which is complementary to fall-out.

The report points out that any of these single measures is inadequate to reflect the performance of a system. Combinations of single measures are required, falling into two groups: twin variable measures and composite measures. For the twin measures one of each of the single measures is taken and a comparison made between them by observing the relative changes in the two values, but retaining each value as a separate entity. The changes in the values result from searching at series of coordination levels.

In the report it is stated that the two major pairs of simple measures are recall with precision and recall with fall-out. Either of these twin measures is satisfactorily for presenting the performance of systems where the generality number is held constant.

In a large number of situations arising in the project, comparison is made between various systems, where everything is being held constant with one exception such as, for instance, the index language. In these circumstances the generality number remains constant and therefore the fall-out measure does not contribute to the presentation of the result.

Composite measures present some compressed and simplified combination of twin variable measures. Perhaps the simplest composite measure suggested is the sum of the recall and precision ratios. The composite measures can be described as linear or non-linear dependent on whether their scale of values varies in a linear or in a non-linear fashion, when recall, precision or fall-out are varied, and the display of the values on twin variable plots results in straight lines or curves.

The report considers the composite measures as inadequate when variables in systems are examined. They may be useful when a single cut-off point is chosen.

AVERAGING SETS OF RESULTS

The results of individual questions will vary considerably, and some idea of the magnitude of this variation can be gained from plots of individual recall and precision ratios for different coordination levels.

To be able to represent generic statements on performance of systems the figures from a set of questions must be averaged in some way. This can be done by taking averages of numbers of documents retrieved or of the ratios calculated for the individual searches.

In the present case the average of numbers method was taken to obtain a comparative ease of calculation. It is stated that the really important matter in any test is to know which method is being used and to use it consistently in all situations. In addition there was the problem of having results for questions which were greatly different in respect of numbers of starting terms and also of retrieving terms. The results are given for use of starting term coordination levels. In this case certain questions having less starting terms than required for the coordination level applied did not contribute. The totalling was simple adding up the results of each question at the various coordination levels. However, a simple method was found of obtaining a ranked output and the majority of the results have been recalculated by the document output cut-off method.

The main presentation of the test results has been on the basis of coordination level cut-offs. The main problem therewith is, that the questions are heterogeneous in having different numbers of starting terms and matching terms.

When making a search for a question at different coordination levels the result will be that at level 1, that is taking all the documents having at least one search term, the best recall is obtained, but that the relevant documents come out together with non-relevant documents. For each subsequent higher level smaller numbers of relevant and non-relevant documents will show up and from a certain level onwards nothing will be retrieved at all. The retrieved documents can now be ranked, giving rank 1 to the first relevant document that shows up without any non-relevant document. The further rank numbers are calculated by means of the formula: $R = X + (n-Y)\frac{(x+1)}{(y+1)}$

R = the rank number determined for the nth relevant document,

X = the number of all documents retrieved at the next higher coordination level,

Y = the number of relevant documents retrieved at the next higher coordination level,

x = the additional number of documents at the nth coordination level,

y = the additional number of relevant documents at the nth coordination level.

The documents are subdivided into a number of groups. For the collection of 200 documents there were 17 groups. The cut-offs used were 1, 2, 3, 4, 5, 6-7, 8-10, 11-15, 16-20, 21-30, 31-50, 51-75, 76-100, and so on. Since the result of the calculation is not always a whole number the following rules have been adopted:

R is taken to the nearest whole number, but if it shows a half, it is taken to the lower whole number for odd numbered questions and to the higher whole number for even numbered questions.

When for a set of questions all the rankings have been calculated for the searches with a single index language, the results are entered on a score sheet. From the score sheet the total number of relevant documents retrieved at each of the cut-off levels can now be obtained, and the recall calculated. The precision ratio is calculated on the basis of the maximum number of documents which could have been retrieved, that is for rank 1 the number of questions and for the higher ranks multiples thereof, obtained by multiplying with the highest number of documents taken for the cut-off.

The recall and precision figures can be plotted on a graph. There is, however, a difference with the graphs obtained with coordination level cut-offs. There, the precision is calculated on the basis of retrieved documents. With document output cut-offs the precision is, however, calculated on the basis of multiples of the number of questions. For each rank recall and precision are in linear relationship. With the coordination level cut-off there is no direct relationship between the various cut-offs. Further it is possible to calculate an average recall ratio, which has been called "normalised recall".

For tests of different systems using the same document/question sets the cut-offs can be applied with equal consistency.

Having developed the simulated ranking method and the method for obtaining normalised recall, the procedure was used for the four main groups of index languages, to wit:

1) eight single term languages,
2) fifteen concept languages,
3) six controlled languages,
4) titles and abstracts.

In a table the index languages are arranged into an order based on the normalised recall. The highest score is obtained by index language I.3.a. (single terms, word forms).

The results, which have been based on the average of numbers have been recalculated by the average of ratios. When placed in order, it can be seen that this order is virtually unchanged from that obtained with the average of numbers.

Also a comparison is made, based on document output cut-off for collections of 1400 and 200 documents with index language I.1.a. and 42 questions. The larger collection shows a smaller generality number.

It is stated that this adversely effects the performance.

With the document output cut-off method the relevance of the documents can be taken into account by giving each document a weighting related to its relevance grading, e.g. a score of 4 to documents rated relevance 1, a score of 3 for relevance 2, a score of 2 for relevance 3 and a score of 1 for relevance 4. Also another weighting was tested.

In total this procedure was done for six languages, but no differences were found when comparing the performance. No particular value of this weighting could be established.

It is emphasised in the report that the normalised recall ratio only has meaning within the context of the manner in which it has been calculated. In this particular case it was by averaging the results of seventeen cut-off groups.

DISCUSSION

The extract of the Cranfield reports given on the previous pages does not comprise all the subjects discussed in the reports. The main aim of the extract is to make the present report self-supporting in respect of the main subject of the Cranfield project. However, in this discussion it will be necessary to refer to the data in the Cranfield reports themselves.

Even when spending only little time on the Cranfield reports it is possible to get an opinion on the work done. From the summary it can already be taken that single terms in natural language give the best performance. However, the words follow "within the environment of this test" and "all results have to be considered within the context of the experimental environment". The warning is, however, tempered by the remark that the characteristics of the subject field will not have made it unique. The conclusions should apparently have some more general importance.

When seeing the enormous amount of work that has been performed in the Cranfield project, one would like to get more out of it than the simple conclusion on single terms in natural language. But then a detailed study becomes necessary. When making a detailed and careful study of the reports it is not easy to reach a point where one has the feeling of having got all out of it. The main task in the present case

was to examine the validity of the conclusions.

I have come to the contradictory result that, although there are details and even quite general points which in a certain way and partly even basically affect the conclusions and the methods by which these have been obtained in Cranfield, the conclusions have a certain value, in particular, within the environment of the project and probably even beyond these boundaries, for example, for another subject. The following remarks may clarify this result of my study and my final conclusion that information cannot be understood within the boundaries set by science and inherent to scientific methods.

1. The mathematical approach.

The performance of the various systems, which have been investigated is indicated by means of a number of ratios, expressed in percentages. This can, however, not always validly be done.

To illustrate this point, reference may be made to the first conclusion expressed in the report, namely, that there is an inverse relationship between recall and precision (within a single system, assuming that the sequence of sub-searches for a particular question is made in the logical order of expected decreasing precision and the requirements are those stated in the question and if the results of a number of different searches are averaged).

There are many examples in the report where recall and precision ratios are not in accordance with this rule, but show a parallel relationship, see e.g. figures 4.103T, 4.140T, 4.303T and many other tables.

The deviation from the rule is always at the higher coordination levels, where only a few documents are retrieved. The precision ratio often becomes even zero, when only non-relevant documents are retrieved and the recall ratio is also zero. The calculations have apparently been pushed forward to a stage where they are loosing meaning.

The mathematical approach is also difficult to follow with the document ranking method. In this method not only retrieved relevant documents are taken up in the performance figures, but also non-retrieved relevant documents and even with recall ratios of 100%. Thus the non-retrieved documents improve the normalised recall.

The precision ratio in the final stage becomes a factor of the generality number, which becomes, in particular, clear for the higher rank number groups. When looking e.g. at figure 5.3T we note for the group of documents numbered 176-200 with a recall ratio of 100%, which

means that 198 documents are retrieved, a precision ratio of 2, which is calculated as follows: $\frac{198 \times 100}{200 \times 42}$. The generality number for a set of 42 questions with a total of 198 pertinent documents and a collection of 200 documents is: $\frac{1000 \times (a+c)}{N} = \frac{1000 \times 198}{42 \times 200}$ which is, with a factor 10, similar to the precision ratio.

It is stated on page 262 of the report that "with the document output cut-off method, recall and precision are, as explained earlier, completely interdependent". When looking for the earlier explanation we find on page 200 that for the various cut-offs recall and precision ratios are interdependent. In the plots which show the interrelationship of recall and precision ratios, for each document output cut-off a straight line can be drawn radiating from the point of origin, which indicates the said interrelationship relating to the particular cut-off. The various precision ratios which are given in the report may have been calculated or taken from plots. For the systems which are being compared the precision ratios differ when the recall ratios differ and only where with different systems for a same cut-off the recall ratios are equal, also the precision ratios are equal to each other.

When in the report the averages of the recall ratios, that is the normalised recall, are compared the precision ratios are omitted. It is not clear why. It is true that if for a certain document output cut-off the recall ratio has been determined, the precision ratio can be found by means of the straight lines in the plots. The precision ratios, therefore, do not contribute independent measures. But there would appear to be no relationship between normalised recall and an average cut-off or an average precision ratio. Figure 5.11T shows e.g. that the sum of the precision ratios is:

368	for	normalised	recall	65.00
375	-	-	-	65.23
379	-	-	-	65.82
352	-	-	-	63.05
361	-	-	-	64.47
354	-	-	-	64.65 (corr.)
340	-	-	-	64.41
293	-	-	-	61.17

The irregularities in the precision ratios mean that the average cut-off is different for the various systems. Normalised recall without an average cut-off or a precision ratio gives incomplete information.

In the Cranfield project great attention is paid to the fallout, the ratio of retrieved non-relevant and the total non-relevant documents. This ratio (F) is of importance together with recall (R), precision (P) and the generality number (G) to recalculate figures to make comparisons possible, since F, R, P and G are interdependent, so that where 3 of them are known the fourth value can be calculated.

In case of a very general question, there will be found many documents that are relevant. Consequently it will be easy to obtain with a high recall ratio a high precision ratio. With a more specific question these ratios may be lower, in particular, the precision ratio, since only a small number of documents will be considered relevant. For comparing systems it is of importance to bring the figures obtained with searches down to a same generality number. This can be achieved by the fact that the four measures R, P, F and G are interdependent. Thus for a given R, P and G the F can be calculated. With this F and the G which it is desired to use in the comparison, for any R the P can be calculated.

Tables are provided for F at R's of 5, 10, etc. up to 100 and P's of 0.5, 1, 2, 3, 4, 5, 7.5, 10, 12.5, 15, 20, etc. up to 100 and that for G's of 1.0, 2.0, 3.0, 4.0, etc. to 10.0, 12.5, 15.0, etc. to 30.0, 35.0, 40.0, 45.0 and 50.0.

These calculations are for the coordination level cut-off procedure. It is therefore quite possible that figures are being compared which have been obtained at different coordination levels. This is undesirable in view of the different situations at the various levels caused by differences in frequencies of use of terms in the system and by differences in existing associations of terms.

Since the calculations have not been used in arriving at the conclusions of the report the matter will not be further investigated. The problem how to compare systems used under different conditions, however, remains. It is a basic problem and it may be discussed in the next section.

2. The comparison of retrieval systems.

As explained in the report, we can distinguish in a search test four entities: a, b, c and d.

a = retrieved and relevant,
b = retrieved and non-relevant,
c = not retrieved but relevant,
d = not retrieved and non-relevant.

In exploiting a system we are among others interested in respect of the performance in the R, which is $\frac{a}{(a+c)} \cdot 100$ and in the P, which is $\frac{a}{(a+b)} \cdot 100$

The R depends on the indexing, whether this has been done by means of terms which are used in formulating questions. The P indicates whether the search terms do not only relate to the subject of the question, but also to other subjects. In other words the P indicates whether the search terms are specific or general, whether they have many associations with other terms in the system or not. This generality aspect finds its expression not only in the P, but better in the fallout, since thereby the relationship to the whole collection comes into the picture. P and F have an inverse relationship. A higher P corresponds with a lower F.

These two aspects of an information system: the quality of the indexing and the nature of the search terms in relation to the system are highly independent. We have to do with two things which do not have any logical or mathematical connection. This only appears when certain conditions are kept constant, such as the type of questions, the size and nature of the collection, the index language, the way of indexing and the way of searching.

In considering only the quantities a, b, c and d the coordination level at which these quantities are obtained is left out of consideration. It cannot be said that this feature is kept constant. It is just left out of consideration. For characterizing the measurements 5 aspects have to be taken into consideration.

The performance of an information system has consequently something that cannot be described when only determining the quantities a, b, c and d. In fact it is even not possible to show diagrammatically the interrelationships of the four quantities. In the Cranfield project a+c and b+d are kept constant, whereas the coordination level is left out in the presentation in the various plots, which in fact only indicate the relation between a and b in the complex forms of the ratios of recall and precision, with the exception of figure 3.21TP, where the coordination levels have been marked.

The situation that we cannot describe a subject in full is met in many cases. If we consider the three major aspects of the human environment: matter, energy and information, we find in this sequence an increase in the difficulties of describing and formalizing them.

A similar problem has been recognized in a special way by S.Kiss, who states in his book "Structures of logic" on page 6, that "an intelligent structural discussion of the biological world" will only be made possible through a 4-class logic which would overcome "the limitations which the exclusive use of 2-class logic has so far imposed upon mathematical thought". S.Kiss writes further that "the Greek scientific civilization was essentially geometrical, and therefore could be readily developed by a 2-class logic, whereas the European civilization of the last four hundred years might be termed physical, because it mainly developed analysis and other branches of mathematics essential for an intelligent structural discussion of the physical phenomena ... a complete solution of the physical problems has not been achieved, a fact attributable to the reliance on 2-class logic to the complete exclusion of a 4-class logic".

We have come in our present days to a new era that is characterized by information. The properties of matter can be described in a three dimensional space. To describe energy we need a further dimension and we may perhaps say that for describing information we still need a further dimension. By recognizing this situation it becomes clear why so many efforts have been made in vain to bring order in structures for, or schemes of information systems.

By arriving at the conclusion that information systems cannot be fully understood in their totallity the interest in the results of the Cranfield project changes. It would not seem to be possible to rationally compare different systems, but only aspects thereof if certain conditions are kept constant.

The conclusion that single terms in natural language give the best results has only a limited value. The conclusion can be overturned by taking certain measures or by looking for other aspects, such as the time involved to obtain a result, which will greatly vary with the level of coordination required to obtain a certain result. From the Cranfield data the conclusion can also be drawn that concept indexing is to be preferred. Not only can concept indexing give high precision when accepting a recall ratio of say about 30%, but such a result can be obtained at a low level of coordination, which means with little effort.

In view of the fact that the results of the preliminary tests were similar to those obtained with the larger collections it could be argued, but it would not be fully justified to say, that the conclusions of the Cranfield project could have been arrived at with much less effort.

3. Importance of limited scope of conclusions

The fact that it cannot be claimed that the conclusions of the Cranfield report are valid for other conditions than those of the environment of the test may on first sight be disappointing. The deduction from this fact that no general measure for comparing and evaluating information systems has emerged from the investigation is of immense importance and can bring a number of existing misconceptions back to reality. Thus, nobody can validly claim in future that this or that system is better than another one. It can only be said that under certain conditions one system may prove to be more suitable.

4. The Index language

The single terms resulting from the "concept indexing" of the collection of 1400 documents form a vocabulary with in total 3094 terms. The frequency of use of these terms proved to be consistent with the well-known Zipf distribution of words according to their frequency of use in natural language texts.

The first ten terms ranked by usage are:

Flow	used 942	times	Theory	used 400	times
Pressure	" 720	"	Velocity	"	"
Boundary	" 512	"	Supersonic	" 360	"
Layer	" 512	"	Mach	" 344	"
Distribution	" 442	"	Equation	" 312	"

1169 terms were used only once.

The index terms are consequently very different as regards their retrieval power. The word "flow" alone, when used in a search will retrieve 942 documents. A term used only once, will when used in a search retrieve 1 document, and when coordinated with "flow" also at most 1 document will be retrieved. When coordinating, however, "flow" and "equation", there will probably be retrieved:
942/1400 x 312/1400 x 1400 or about 210 documents.

Since it is a basic requirement for a uniform output of an information retrieval system that the terms have about the same frequency of use, it is clear that the Zipf-distribution of the terms causes great differences in performance.

A more exhaustive indexing will bring more specific, that means also less frequently used terms; consequently any change in depth of analysis in the indexing stage will change the nature of the system. Moreover, attention should be paid to the extent to which words are associated. If two search terms are strongly associated, which means that they have very often been used together, then there will be retrieved more documents than will be found by way of the probability calculation just referred to. The inner structure of the vocabulary as regards association of terms can be established, but in case the structure should be comprehensive and not just for a small number of terms, the use of a computer to establish such structure is to be recommended.

5. Averaging groups of results

The problems arising in connection with averaging results of a number of search questions have well been recognized in the report.

The various documents of the collections show all sorts of variation in respect of the presence of search terms. In the tests the results were collected for the different coordination levels. For example, if there were 7 starting terms, then at coordination level 3, all documents that have any 3 terms out of the 7 terms would be taken as retrieved. By this method all qualitative aspects are getting lost. Thus the influence of the Zipf distribution will disappear, since among the search terms there will be a number of frequent terms and a number of non-frequent terms. The frequent terms will cause the retrieval of most of the documents.

For comparing the performance of different systems this averaging of results may be acceptable. For getting an insight in the influence of the various factors a subdivision in various types of searches may be of interest. This has been done to a certain extent in the Cranfield Project. But there seems to be more room for further investigations with the material that is now available.

When averaging results obtained with the coordination level cut-off, there will, at higher levels, be documents that no longer contribute in view of the fact that in the indexing less terms have been used than are required for the searching. With the document output cut-off a comparable feature turns up. For all questions all the retrieved relevant documents are ranked. For calculating the precision it is supposed that when no relevant document is retrieved, that there is then a non-relevant document. The various questions show for the last docu-

ment that is ranked different rank numbers. In other words the part of the documents in which the relevant documents are found is different for the various questions. In case this part is small the result of the search contributes more to the normalised recall than when the part is larger. It would be possible to group the questions on the basis of the parts of the collection in which the relevant documents are ranked, just as with the coordination level cut-off the questions can be grouped in respect of the number of starting terms or retrieving terms. This grouping of results may have an influence on the comparison of systems.

With the ranking procedure the normalised recall is improved by questions showing lower rank numbers or in other words having smaller parts of the collection in which the relevant documents are ranked. This seems to be an interesting aspect of the ranking method. In the coordination level cut-off it should be the level that has to be taken into consideration. This aspect may be further discussed.

6. The levels in the coordination level cut-off

All results of the tests of the Cranfield project have been obtained at distinct coordination levels. These disappear, however, in the presentation of the results in plots. Thereby the effect of the application of recall devices becomes less clear.

It is obvious that grouping of index terms must improve recall. Thus, grouping of different word forms, synonyms, etc., must improve recall, and as is also obvious, at the cost of a lower precision. This is readily seen when plotting the results for recall ratios versus corresponding coordination levels. The precision ratios can be plotted independently versus the coordination levels.

If we compare, for example the recall ratios for a few index-languages I for 1400 documents, 221 questions and 1590 relevant documents given in figures 4.10 (1-4) the following table can be made:

Language:		I.1.a	I.2.a	I.3.a	I.5.a	I.6.a
level	1	95.0	95.2	96.4	97.4	97.9
	2	80.7	82.6	84.2	88.4	89.9
	3	59.5	61.7	64.0	70.5	73.3
	4	38.1	40.5	42.6	50.4	53.3
	5	19.7	22.3	23.5	29.9	31.6
	6	9.7	10.6	12.1	16.7	18.6
	7	4.7	5.0	6.0	8.2	10.1

The table shows: The larger the classes of index terms the better recall at each of the coordination levels.

When looking at the normalised recall, the effect of grouping is no longer visible. Also in the ranking procedure the coordination levels are not taken into account. Thus in one case a rank 1 document may have been obtained at coordination level 4 and in another case at coordination level 9.

It would seem that plotting recall and precision ratios against coordination levels is a simple and effective way for comparing information systems.

A composite measure could be obtained by plotting the product of recall and precision ratios versus coordination levels. Then optima will show themselves, which very well represent performance optimas for practical purposes.

7. Practical results of the Cranfield Research

The Cranfield Research has been done in a laboratory sphere. The question is what use can be made in practice of the work done.

In a practical situation it is impossible to determine how many documents are relevant in respect of a certain question. An estimate can be obtained by taking a sample and to investigate for the sample the relevance of the documents retrieved in a search in respect of the question and to calculate recall and precision ratios. The sample has to be taken at random and it must yield figures of some magnitude, which will only be the case if the questions are relatively broad.

Although in the Cranfield Project collections of different sizes have been used, these collections do not fulfill the requirement that the small collections are randomly taken from the larger one. On the contrary all the relevant documents in the small collection are the same as those in the large collection. The research work done does not give information in respect of the question to what extent the sampling method gives practical results.

A more attractive method would seem to be to accept figures obtained in the Cranfield project for recall ratios at various coordination levels. For an operational system one could accept e.g. that at a coordination level of 3 the recall ratio is about 50%. However, in the Cranfield project coordination level 3 means that out of e.g. 7-10 search terms all possible combinations are taken, but are counted only once independent of which terms match. This is inherent to the scan-column index technique used in the project, which is not easy to imitate in operational systems using computers, machine punched cards or manual punched cards. One way is when a document is found

more than once to count it only once. On the other hand figures can be obtained on the matches of the various combinations of search terms, which was not the case in the Cranfield Project. When using this method for investigating the performance of a system only precision ratios can be determined, whereas the recall ratio taken from the Cranfield results gives an indication of the part of relevant documents that has been retrieved.

The methods used in the Cranfield Project give a general idea of the performance of a system. What one is really interested in, is, having a certain question, what the system will yield. Since non-relevant documents are unwanted the first approach for a search will be to get information - in the form of references to documents - with high precision. In case no suitable documents have been found the search can be broadened by using a lower coordination level. With 3 or 4 search terms it will still be possible to search for all combinations of terms. With 5 search terms 31 searches would be required, which no longer is warranted.

The method of searching in an individual case depends greatly on the insight of the searcher. When comparing, for a certain question the precision ratios when having made searches with e.g. 3 out of 4 search terms, it can be seen which terms have many associations with terms used for other subjects than the subject for which the search is done. The precision ratio will in such a case be lower. If the search question is a broad one the documents found with low precision may be of interest. For a very specific question those term combinations yielding a high precision may be more useful. All these aspects are closely related with the frequencies of use of the search terms. When knowing these frequencies the most suitable searches can be determined by means of probability calculations. This aspect has not found the attention which it deserves. This can be well understood, since the aim was to obtain average results for large sets of questions. In practice one would like to know what to do in a specific case. The material for further studies is available and further use thereof can undoubtedly be made.